Differing World Views in Modeling

Deterministic vs Stochastic

Mary J. Camp

Biometrical Consulting Service
USDA, ARS, Beltsville

Begin at the Beginning

- Observation – a record obtained by an act of recognizing and noting a fact or occurrence often the outcomes of an experiment, investigation, or survey and measuring with instruments.
- Data – a collection of observations. Factual information (as measurements) used as a basis for reasoning, discussion, or calculation.
- What to do with data? Investigate how it came about, what caused it, manipulate conditions to produce it, use it to make predictions.
- To do the above to data usually means – Model It

World Views in Modeling

- How data is modeled and the purpose of modeling will depend on the modeler's world view.
- Deterministic Model (Functional Model)
- Stochastic Model (Statistical Model)
- Classical Statistical Model (General Linear Model)
  – actually a subset of the Stochastic Model

Classical Statistical Modeling

- An observation is thought of as being composed of three parts: a part due to the average of all observations in the population, a part due to manipulation, the level of an applied factor(s), i.e., a treatment(s), and a part due the unique properties of that particular observation in the population

$$Y_{ij} = \mu + \tau_j + \epsilon_{ij}$$

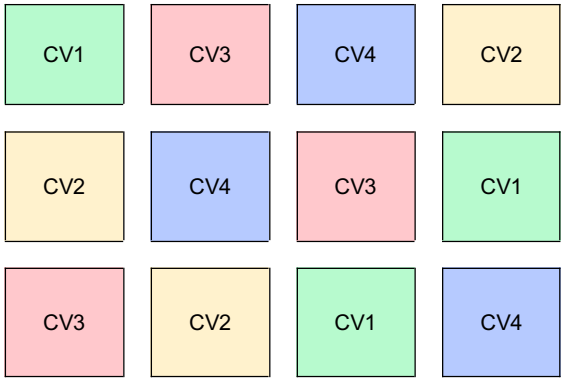
- Rearranging shows that the deviation of an observation from the overall mean is then due to the effect of its factor level, the treatment effect, and its other properties, the error

$$Y_{ij} - \mu = \tau_j + \epsilon_{ij}$$

Assumptions of the Model

- The observations are independent. Measuring or observing one does not affect the measurement or observation of another.
- The error is a sample from a probability distribution. Often in modeling this is a normal distribution, also known as the bell-shaped curve.
- The errors for the observations come from the same probability distribution.

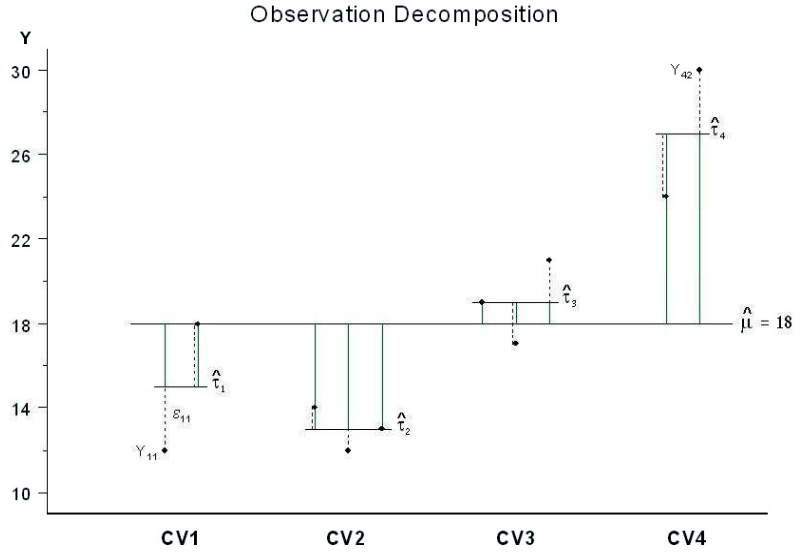
Example 1: High Tunnel Tomato Yield



	Cultivar				
Plot	1	2	3	4	
1	12	14	19	24	
2	18	12	17	30	
3	—	13	21	—	
					Total
Total	30	39	57	54	180
Mean	15	13	19	27	18
Number of Plots	2	3	3	2	10

An observation, the yield for a plot, is viewed as being composed of the average yield of tomatoes in the high tunnel plus an effect due to the cultivar on the plot and an effect due to the individual differences intrinsic to each plot.

$$Y_{ij} = \mu + \tau_j + \epsilon_{ij} \qquad i = 1\ldots3, \quad j = 1\ldots4$$



Determining if Treatment is Important

- For each observation the square of the distance of the estimated treatment effect from the estimated overall mean is calculated. In the above plot would be squaring and summing the solid green lines. The sum of these distances is known as the *sum of squares treatment*.
- The squared distance of the each observation from the estimated treatment effect is calculated. In the above plot this would be squaring and summing the dotted red lines. The sum of these distances is known as the *sum of squares error*.
- We look at a ratio of the average sum of squares treatment and the average sum of squares error.
- If the ratio is large enough, then we judge that the treatment has an important effect in understanding the differences between the means of the treatment levels.

High Tunnel Tomato Yield Analysis

- The average sum of squares treatment, i.e., the cultivar effect: 258/3 = 86
- The average sum of squares error, i.e., the mean square error: 46/6 = 7.67
- The ratio is: 86/7.67 = 11.21
- Under the assumptions of the model and that the probability distribution for the error is the normal distribution, 11.21 is large enough. The probability of obtaining a ratio this large if the average cultivar yields were not different is only .007. The conclusion is that the cultivar is important in explaining the differences in the average tomato yields.

	Cultivar 1	Cultivar 2	Cultivar 3	Cultivar 4
Average Yield	15	13	19	27

Deterministic (Functional) Model

- Mathematical function(s) is used to model a process, usually chemical or physical.
- Observations or predictions are the results of how the inputs interact in the process.
- The model can be very complex however the more complex the model the more inputs, *parameters*, and terms are needed for prediction.
- The model is only as good as the science used to make it. Assumes the process is understood and the data for it can be collected.
- By changing any of the inputs, any of the values of the parameters, new predictions and “What if?” questions can be asked.

Example 2: Return on an Investment

$$F = P(1 + r/m)^{Ym}$$

Where:

F = Future value
P = Present value,
r = Annual rate,
m = Periods/Year,
Y = number of Years

5-Year Return on $1000 at Federal Funds Rate, June 2004 – January 2006

Rate	Return	Rate	Return	Rate	Return	Rate	Return
1.25	1064.46	2.25	1118.95	3.25	1176.19	4.25	1236.30
1.50	1077.83	2.50	1133.00	3.50	1190.94	4.50	1251.80
1.75	1091.37	2.75	1147.22	3.75	1205.88		
2.00	1105.08	3.00	1161.62	4.00	1221.00		

Example 3: Verhulst-Pearl Logistic Growth

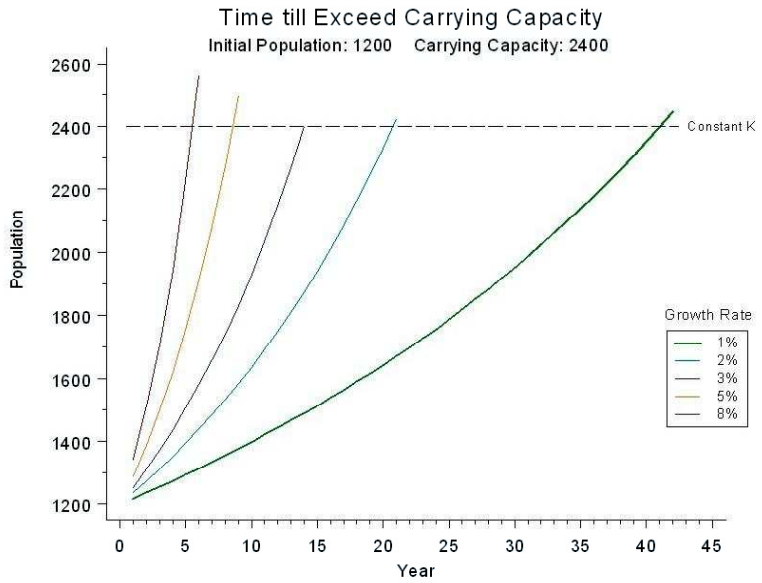
$$N_{t+1} = N_t + rN_t(1 + N_t/K)$$

Where:

N_t is the population size at time t,

r is the growth rate of the population,

K is the carrying capacity of the habitat



Stochastic (Statistical) Model

- Uses mathematical function(s) to model a process.
- At least one model parameter is a random variable described by a probability distribution.
- Ability to reproduce and predict observations are based on patterns of previous data, not necessarily the underlying physical or chemical processes.
- Some view these models as 'black-box' models.

Example 4: Verhulst-Pearl Logistic Growth

$$N_{t+1} = N_t + rN_t(1 + N_t/K_{f(t)})$$

Where:

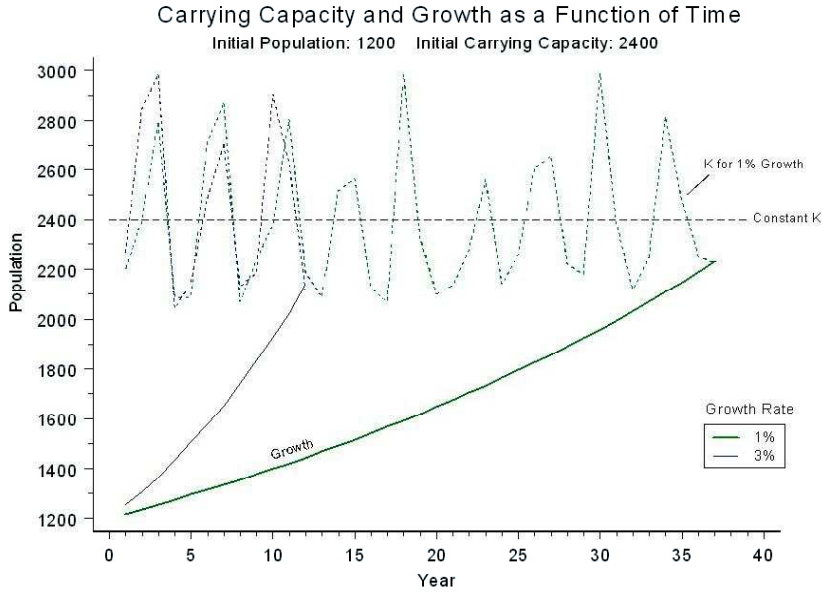
The carrying capacity at time t, $K_{f(t)}$, is the initial carrying capacity multiplied by a random variable, *p*. $K_{f(t)} = K_0 p$

In years 1,4,5,8,9,12... p has an equal chance of taking any value between 0.85 and 0.95.

In years 2,3,6,7,10,11... p has an equal chance of taking any value between 0.95 and 1.25

Mathematically this is

$$\begin{cases} 0.85 \le p \le 0.95 \text{ uniformly distributed when } t = 1,4,5,8,9,12... \\ 0.95 \le p \le 1.25 \text{ uniformly distributed when } t = 2,3,6,7,10,11... \end{cases}$$



Example 5: Number of Marriage Licenses

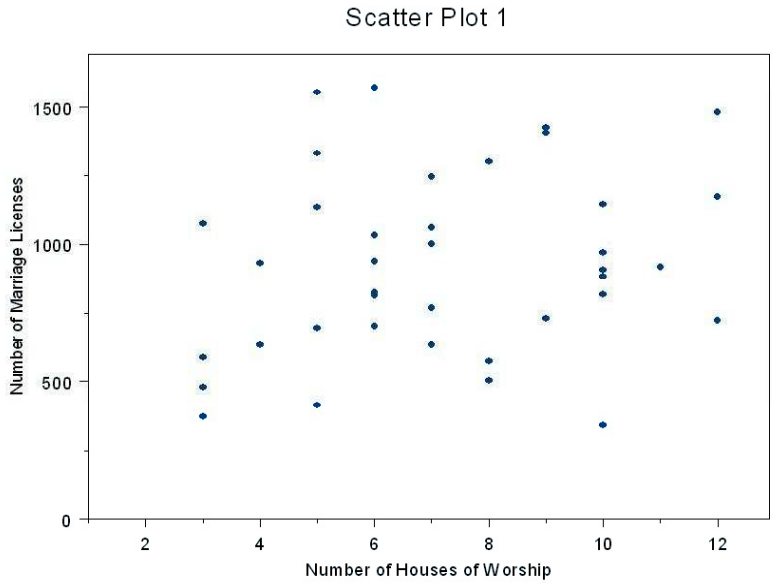
In an attempt to model the number of marriage licenses issued in March from 38 randomly selected county courthouses, a linear regression model was used.

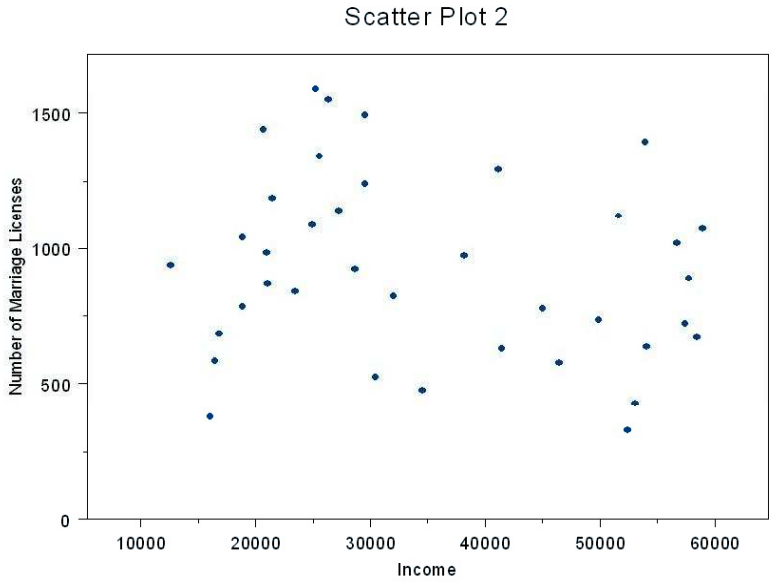
The input variables were:

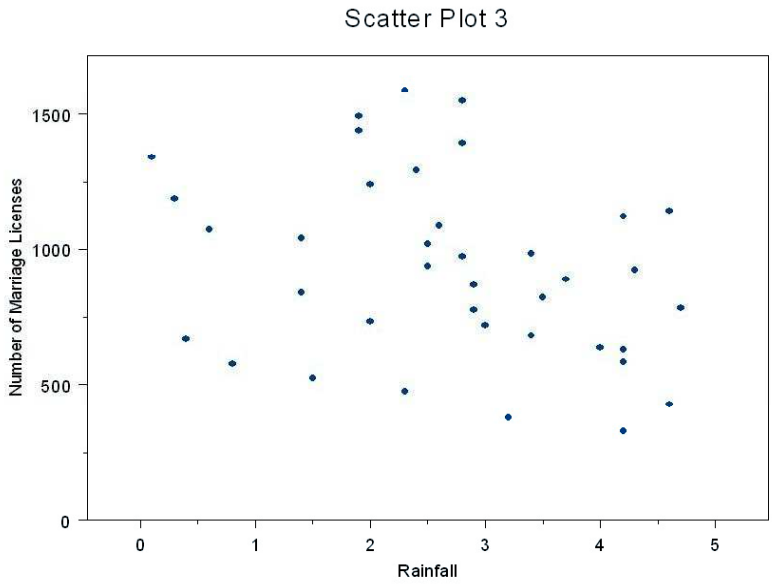
Income	=	average household income in the county.
Liquor	=	number of liquor stores within a 10 block radius of the courthouse.
Rain	=	rainfall in inches for the county in March.
Robins	=	number of robins reported in the county's March bird survey.
TV	=	average number of television sets per household in the county.
Worship	=	number of houses of worship within a 3 mile radius of the courthouse.

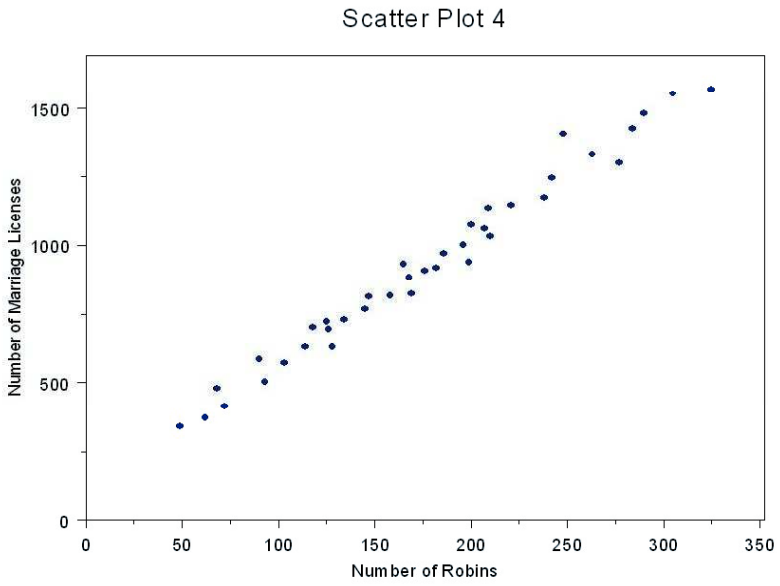
Result

Marriage Licenses = 108.90 + 4.63(Robins)









Comparing Models

- Deterministic and Stochastic models look the same when the deterministic model has model error, a random part.
- Deterministic model with error

$$Q = f_d(P|\Omega) + \epsilon$$

Q is the observation based on a function, $f_d(\cdot)$, of the process P, for the model parameters, Ω, and ϵ is the model error.

- $f_d(P|\Omega)$ is the deterministic component and ϵ is the stochastic part.
- The model error can occur either through mis-specifying the model, i.e., leaving out factors that explain the process, or measurement error.
- Goal is usually to minimize ϵ by some means and focus on the deterministic element.

- Stochastic model with error

$$Q = f_s(P|\Omega) + \nu$$

Q is the observation based on a function, $f_s(\cdot)$, P is the inputs, for the model parameters, Ω, and ν is the model error.

- $f_s(P|\Omega)$ is the deterministic component and ν is the stochastic part.
- The model error occurs because the model is based on mathematical function(s), measurement error and unexplained variability.
- In the stochastic model the deterministic element, $f_s(P|\Omega)$, is derived to insure reproduction of characteristics of Q without regard to underlying physical processes.

- The Stochastic model's weakness is that it is does not necessarily represent observed internal physical laws or processes. Consequentially, the model may not be useful in understanding how the observations occur.

- The Deterministic model's weakness is that it can not reproduce the variance of observed model outputs. As long as the model residuals (observed value – predicted value) are independent of the model inputs

$$Var[Q] = Var[f_d(P|\Omega)] + Var[\epsilon]$$

it will always hold that $Var[f_d(P|\Omega)] < Var[Q]$, unless $Var[\epsilon] = 0$ which means there is no model error.

Relationship to Spatial Modeling

Spatial models comprise two sources of variation:
Large Scale Variation (modeling the mean structure) and
Small Scale Variation (modeling the covariance structure).

Large Scale Variation = Trend

- Involves the entire region of the study or experiment area.
- All points are used equally to predict an observation.
- In a deterministic model this would be the functional part that describes a process, e.g., modeling how fast water flows down a slope.
- In a stochastic (statistical) model this would be the treatments in an analysis of variance, independent variables in a regression, blocks.

Small Scale Variation

- Once large scale variation has been removed, only neighboring points are used to estimate a nearby observation.
- Observations are viewed as being correlated. Observations close together are more correlated than observations further apart. As observations become further apart a distance is reached where the correlation is negligible.
- A perfect deterministic model would have no small scale variation. That these models do have small scale variation is largely a matter of measurement error.
- A statistical model will have small scale variation, since the model is based on mathematical functions and proxy variables that do not fully explain the process, plus it will have measurement error.

Some physiologists will have it that the stomach is a mill; --others, that it is a fermenting vat;--others again that it is a stew-pan;--but in my view of the matter, it is neither a mill, a fermenting vat, nor a stew-pan--but a *stomach,* gentlemen, a *stomach.*

William Hunter 1718–1783

References

J.L. Cisne. How Science Survived: Medieval Mansuscripts' "Demography" and Classic Texts' Extinction. *Science*, **307**:1305–1370 (2005).

J. W. Hayse and I. Hlohowskyj. Comparison of Deterministic and Monte Carlo Analyses for Evaluating Risks to Ecological Receptors With Contaminant Uptake Models. (1998) http://web.ead.anl.gov/jfield/PPT_presentations/ArmyPresentation/index.htm

A. J. Lembo, Jr. Lecture 3: Model Use and Development Spatial Modeling and Analysis. www.css.cornell.edu/courses/620/lecture3.ppt

J. Neter and W. Wasserman. *Applied Linear Statistical Models*. Richard D. Irwin, Inc, Homewood, 1974

O. Schabenberger and C. A. Gotway. *Statistical Methods for Spatial Data Analysis*. Chapman and Hall/CRC Press, Boca Raton, 2005

R. M. Vogel. Stochastic and Deterministic World Views. *Journel of Water Resources Planning and Management*, **125(6)**: 311–313 (1999)